Running head: Alternatives to Traditional Vertical Scales

Un-Distorting Measures of Growth: Alternatives to Traditional Vertical Scales

Joseph A. Martineau

Michigan Department of Education

Office of Educational Assessment & Accountability

Paper presented at the

35th Annual National Conference on Large-Scale Assessment

of the Council of Chief State School Officers

June 19, 2005

Abstract

This article reviews some serious distortions in measures of student growth and in measures of educator effectiveness that arise from the use of traditional vertical scales in growth-based statistical models. For pure growth models, these distortions include identification of growth trajectories that have little resemblance to true growth trajectories, the attribution of effects on growth to effects on initial status and *vice versa*, the identification of false effects on either initial status or growth, failure to detect true effects on either initial status or growth, and the identification of effective interventions as harmful and *vice versa*. For Value-Added Models, the distortions include the mis-estimation of educator effectiveness simply because educators serve students whose growth is occurring outside the range measured well by the test, and the attribution of prior educators' effectiveness to later educators. This article also reviews follow-up work on alternatives to traditional vertical scales. This article concludes that alternatives which have been studied are important to, but insufficient in resolving the distortions. It further concludes that alternatives which have not yet been studied are promising, but will be challenging to implement.

Un-Distorting Measures of Growth: Alternatives to Traditional Vertical Scales

Using Vertical Scales to Measure Growth

Much historical work has examined the validity of traditional vertical scales for use in the measurement of student growth (Angoff, 1971; Bereiter, 1963; Cliff, 1991; Hoover, 1984; Lord, 1963; Phillips & Clarizio, 1988; Reckase, 1989a; Seltzer *et al.*, 1994; Yen, 1985, 1986, 1988; Zwick, 1992). Recent work at the intersection of the growth modeling and psychometrics has identified some difficulties in the measurement of student growth, and in the measurement of educator effects on student growth (Martineau *et al.*, submitted for publication; Reckase, 2004; Schulz *et al.*, 2005).

Reckase (1989a) and Reckase and Martineau (2004) showed that even on assessments that have traditionally been considered unidimensional, multidimensionality plays an important role. They divided the unidimensional estimates of achievement into equally-spaced intervals and calculated the average score on each multiple dimension for each interval. The result was that increases on the unidimensional score scale correlated with comparable increases in *different* subsets of the multiple dimensions of achievement depending upon the portion of the unidimensional scale being analyzed. This indicates that the meaning of scores varies along the length of the score scale. For the measurement of growth, this is a serious violation of assumptions (see Bereiter, 1963; Bryk *et al.*, 1998a; Thum, 2002).

In addition, Martineau (2004, submitted for publication-a, submitted for publication-b, submitted for publication-c) showed that serious distortions in measures of growth and effects on growth arise from the violation of the assumption that content representation does not change across the length of the scale. Martineau showed that these serious distortions include the identification of growth trajectories that have little resemblance to true growth trajectories, the

attribution of effects on growth to effects on initial status and *vice versa*, the identification of false effects on either initial status or growth, failure to detect true effects on either initial status or growth, and the identification of effective interventions as harmful and *vice versa*.

Graphical representation of the distortions in growth

The effects of these distortions can be explained graphically as shown in Figure 1. Figure 1 assumes that the scale within each grade is linear, when this is likely not the case (see Reckase, 1989a; Reckase & Martineau, 2004 for examples of non-linear scales within grades). It also assumes that the transition from grade-5 to grade-6 content over the vertical scale is smooth. Even with these simplifying assumptions, the change in content across grades presents a tremendous challenge, as shown below.

In panel A of Figure 1, the lines represent the unequated grade-5 and grade-6 mathematics scales. The x-axis represents student achievement on number sense, and the y-axis represents student achievement on algebra. Because the unequated grade-5 scale is nearly parallel with the x-axis, the unequated grade-5 scale measures mostly differences in number sense. Because the unequated grade-6 scale is nearly parallel with the y-axis, the unequated grade-6 scale measures mostly differences in algebra (see Schmidt & Houang, 2004 for an empirical rationale for this dramatic change in content from grade 5 to grade 6). The unequated grade-level scales are repeated in panel B.

In panel B of Figure 1, the grade-5 and grade-6 scales are equated to create a single vertical scale covering grades 5 and 6. In order to link the grade-5 and grade-6 scales, the scale has to be bent in the middle to accommodate both unequated scales. When achievement in number sense and algebra are strongly correlated, the “equated” vertical scale can appear to be unidimensional. The vertical scale is repeated in panels C-F.

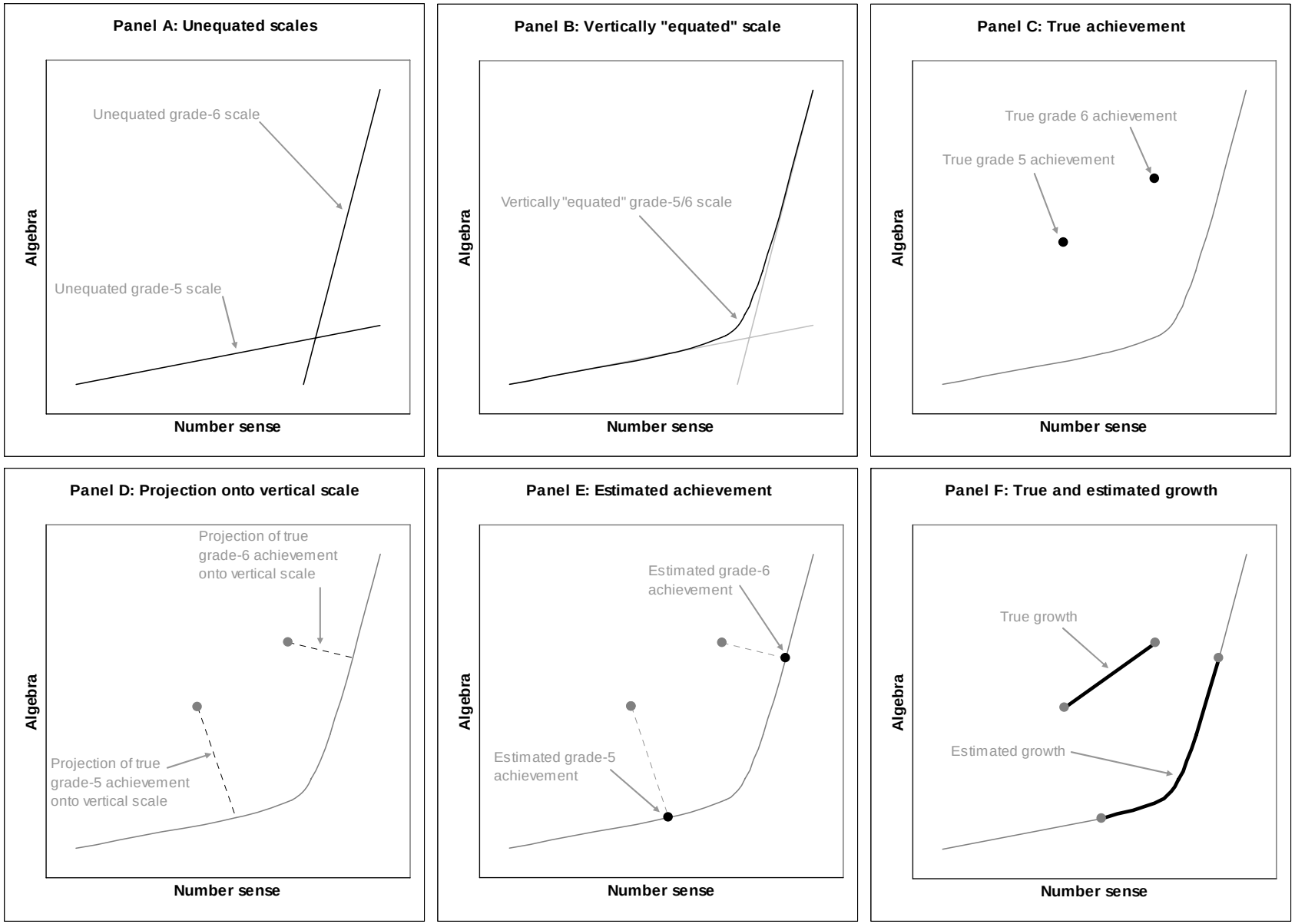

Figure 1. Graphical demonstration of a distortion in growth attributable to vertical scaling.

The solid dots in panel C of Figure 1 represent a given student's true grade-5 and grade-6 achievement. Note that this student's achievement lies outside the range of achievement measured well by either the fifth grade or sixth grade test (the solid dots are relatively far from the lines representing the vertical scale). These dots representing true achievement are repeated in panels C-D.

The dotted lines in panels D and E are the perpendicular lines from the student's true grade-5 and true grade-6 achievement to the vertical scale. Because a student's two-dimensional achievement is reduced to a unidimensional scale, the multiple true achievements are *projected* onto a single score scale (see Reckase & McKinley, 1991). With perfect estimation, the projection will follow the dotted lines shown in panels D and E.

The solid black dots located on the grade-5 and grade-6 scales in panel E represent asymptotic estimates of grade-5 and grade-6 achievement on the vertical scale.

The heavy lines in panel F of Figure 1 represent the true multidimensional growth and perfectly estimated "unidimensional" growth based on the grade-5/6 vertical scale. In the particular illustration in Figure 1, the distortion is that the estimated growth is approximately twice as large as the true growth achieved by this student. In other situations, it could be a severe underestimation of growth. The nature of the distortion varies considerably with differences in true student achievement.

The distortions come about because of the conflict between the requirements of the statistical models and that actual complexity of student test data. In part to address this disconnect, the entire first issue of the 2005 volume of *Applied Measurement in Education* (Cizek, 2005b) is devoted to vertically moderated standard setting. The goal of that issue was improving the quality of vertical scales for measuring growth in accountability applications. In

addition, the lead article in a recent issue of *Journal of Educational Measurement* (Schulz *et al.*, 2005) describes a method of measuring growth that in theory avoids the issue of changing content across grades by grouping test items into theoretically defined content domains and measuring gains across the multiple domains.

Using Vertical Scales to Model Growth-Based Educator Effectiveness

The interest in the measurement of student gains has increased dramatically over the last few years with the advent of Value Added Models (VAM) on the educational policy stage, increasing the need for attention to the disconnect between statistical needs and measurement realities. Examples of this increased interested are that the entire Spring 2004 issue of *Journal of Educational and Behavioral Statistics* (Wainer, 2004), the entire December 2004 issue of *The School Administrator*, and approximately half of the Summer 2003 issue of *Education Next* were devoted to VAM. In addition, a search of the *Education Week* archives on www.edweek.org using the search term ["Value added" or "Value-added"] returned 45 articles from 1 January 2003 to 8 June 2005 inclusive. The central topic of twelve of those 45 articles was VAM, while VAM played a supporting role in the rest. Finally, four prominent national conferences on VAM took place within a month of each other in late 2004[1]. This dramatic increase in interest in VAM covers a broad range of audiences and technical expertise. The assumptions about measurement made in growth-based VAM are the same assumptions made in the measurement of growth.

As with the assumptions of the statistical growth models, many scholars have taken issue with the assumptions of VAM. Ballou (2002), for example, indicates that one of the pitfalls of

---

[1] The four conferences were The University of Maryland Conference on *Value-Added Modeling: Issues with Theory and Application*, College Park Maryland, October 2004; The *CCSSO Brain Trust on Use of Growth Models Based on Student-Level Data in School Accountability*, Washington DC, November 2004; The *CRESST Conference on Value Added Models*, Los Angeles CA, September 2004; The Center For Assessment Conference on *Incorporating Measures of Student Growth Into State Accountability Systems*, Nashua NH, October 2004.

VAM is that the statistical models make unrealistic assumptions about the comparability of vertical scales along the entire length of the scale. Reckase (2004) also expressed uneasiness about the assumptions of VAM concerning the characteristics of the vertical scales they use. Schmidt and Houang (2004) similarly voiced the concern that the assumption that the content on which the scales are based does not change over grades is an unrealistic assumption, and probably biases the results of measures of educator effects on student growth. Braun (2004) listed several additional assumptions of VAM that may be unreasonable.

Popham (1997) cautiously urged the use of VAM for the evaluation of educational effectiveness, after acknowledging that he had been drawn to the idea of evaluation of educators based on student learning "as a moth to the flame" (p. 264). He expressed some cautious optimism that with VAM, some of the myriad difficulties in this venture appeared to be reasonably overcome. However, Popham's view has now changed. At the 2005 meeting of the National Council on Measurement in Education, Popham (2005) served as a discussant to Kingsbury and McCall (2005) who asserted that vertical scales can be created to satisfy the statistical needs of VAM. Popham responded that to make vertical scales sufficient for VAA, the content at each grade level would have to be so vague as to represent intelligence rather than achievement, a weak basis for evaluating instructional effectiveness.

Strong advocates of VAM also acknowledge that to the degree the scales do not meet the assumptions of the statistical models, the models produce biased and/or distorted results (see Bryk *et al.*, 1998b; Sanders as quoted in Schaeffer, 2004; Thum, 2002).

Martineau and Plank (Martineau, 2004, in press; Martineau & Plank, submitted for publication) have shown that serious distortions arise because of the assumption that the content on the tests does not change across grade levels. These distortions include (1) the mis-estimation

of educator effectiveness simply because educators serve students whose growth is occurring outside the range measured well by the test (e.g. this distortion applies to educators of gifted and talented, very high socioeconomic status, very low socioeconomic status, and disadvantaged students), and (2) the attribution of prior educators' effectiveness to later educators.

Martineau (in press) showed that the attribution of prior educators' effectiveness to later educators causes the reliability of *population parameters* of growth-based VAM to be insufficient to support high-stakes use. Martineau and Plank (submitted for publication) showed through effect size calculations that the effect of misattribution of previous educators' effects on mis-estimation of educator effectiveness is practically significant except in the most unrealistic of assessment circumstances.

The use of traditional vertical scales for measurement of growth and for VAM cannot be supported, and if growth models/VAM are to implemented in high-stakes situations, an alternative to traditional vertical scales must be found.

Graphical representation of the distortions in VAM

Distortions from to student achievement lying outside the range measured well by the test

The effects of the first type of distortion (mis-estimation of effectiveness because student abilities lie outside the range measured well by the test) can be explained graphically as shown in Figure 2. Figure 2 makes the same simplifying assumptions as in Figure 1. Again, even with these simplifying assumptions, the change in content across grades presents a tremendous challenge, as shown below.

Panel A of Figure 2 shows the vertically "equated" grade-5/6 "unidimensional" score scale. The solid squares in panel B show true average statewide achievement scores. The solid dots in panel C show true average achievement scores for a given school (school *X*). Panel D

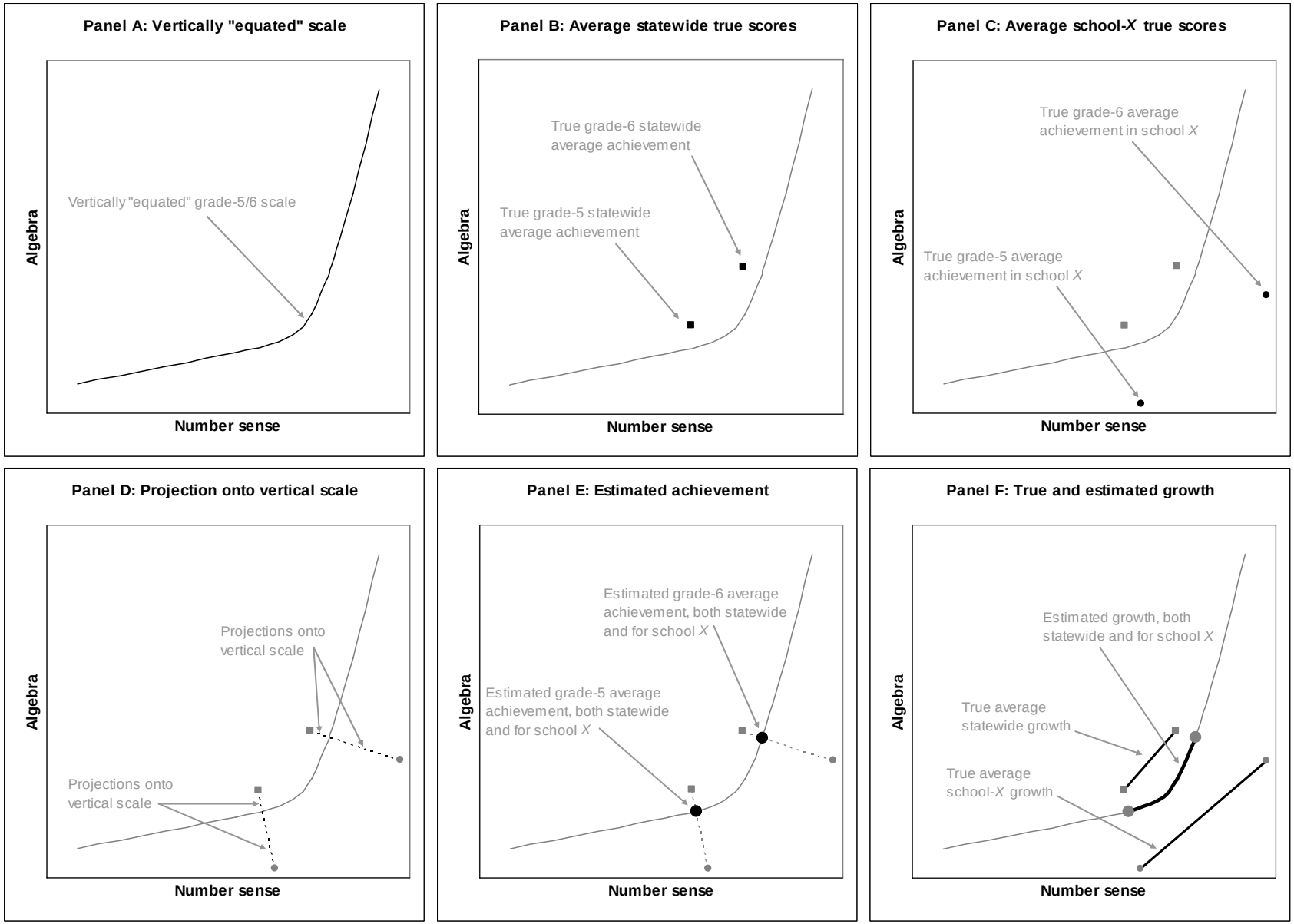

Figure 2. Graphical demonstration of a distortion in VAM attributable to vertical scaling

shows the projection of the average scores in panels B and C onto the vertical scale. Note that both the statewide and school-*X* averages project onto the same location on the vertical scale. Panel E shows that the estimated statewide averages are the same as the estimated school-*X* averages even though they are quite different in reality. Finally, panel F shows that although the true average growth of school *X* is much larger than the true average statewide growth, the estimated growth is the same for the entire state and for school *X*.

In a statewide VAM, school *X* would be found to be of average effectiveness, when in fact it is obvious from Figure 2 that children in school *X* gained far more than the average growth observed across the entire state. School *X* should have been identified as highly effective when it was identified as only average. In other situations, the distortions can result in highly ineffective schools being identified as average. Both incorrect designations as average can have tremendous educational implications for individual students.

Distortions from contamination with previous educators' effectiveness[2]

The second type of distortion (contamination of later educators' effectiveness estimates with previous educators' effectiveness) is more difficult to represent graphically. However, this type of distortion can be represented in tabular format. In order to understand the table, it must first be explained this type of distortions is affected by five variables:

1. The largest allowable proportion of educators affected by a distortion (the smaller the allowable proportion affected, the larger the effect size of the distortion for that proportion of educators),
2. The number of previous grades in the analysis (the more previous grades, the larger the effect size of the distortion),

[2] The numbered list and table in this section were copied from Martineau and Plank (submitted for publication).

3. The change in content representation across grades (the larger the shift in content representation, the larger the effect size of the distortion),
4. The within-educator correlations between value added by student gains on the multiple constructs included within a given assessment instrument (e.g. the larger the correlation between value added to number sense and value added to algebra, the smaller the effect size of the distortion), and
5. The within-grade balance of the representation of the multiple constructs in students' achievement scores (The more equal the content coverage within each grade, the larger the effect size of the distortion. This unintuitive result is deduced from the formula for the effect size in Martineau & Plank (submitted for publication)).

Table 1 shows the effect sizes of these distortions for different values of each of the five variables. Table 1 shows that unless the value added by educators on the multiple dimensions is highly correlated (greater than 0.7), the content mix on the test is unbalanced, the change in representation from one grade to the next is small, the number of prior grades included in the analysis is only one or two, and the acceptable proportion of educators whose estimates will be significantly distorted is large, the effects are of practical significance. This indicates that the effects of the distortions have practical effects on most educators included in a VAM analysis.

Alternatives to Traditional Vertical Scales

Vertically moderated standard setting

Vertically moderated standard setting (VMSS) has been put forward as an *alternative* to vertical scaling (see introduction to Huynh & Schneider in Cizek, 2005a; Huynh & Schneider, 2005). VMSS is a good beginning on addressing the problems identified by Martineau (see above). However, it does not resolve the distortions because movement across vertically

Table 1. *Effect Sizes of Distortions Arising from Violating Value Added Models' Statistical Need for Unchanging Content Across Grades.*

Construct Representation[a]		# of prior grades in the analysis	Minimum Effect Size for…																						
			10% of educators, with[b] ρ_{a_1,a_2} =…							25% of educators, with ρ_{a_1,a_2} =…							50% of educators, with ρ_{a_1,a_2} =…								
Status	Change		0.0	0.1	0.3	0.5	0.7	0.9	1.0	0.0	0.1	0.3	0.5	0.7	0.9	1.0	0.0	0.1	0.3	0.5	0.7	0.9	1.0
0.2 (unbalanced)	0.1 (small)	1	0.28	0.26	0.22	0.18	0.13	0.07	0.00	0.20	0.18	0.15	0.13	0.09	0.05	0.00	0.12	0.11	0.09	0.07	0.05	0.03	0.00
		2	0.56	0.52	0.44	0.36	0.27	0.15	0.00	0.39	0.37	0.31	0.25	0.19	0.10	0.00	0.23	0.21	0.18	0.15	0.11	0.06	0.00
		3	0.85	0.78	0.66	0.54	0.40	0.22	0.00	0.59	0.55	0.46	0.38	0.28	0.16	0.00	0.35	0.32	0.27	0.22	0.16	0.09	0.00
		4	1.13	1.05	0.88	0.72	0.54	0.30	0.00	0.79	0.73	0.62	0.50	0.37	0.21	0.00	0.46	0.43	0.36	0.29	0.22	0.12	0.00
		5	1.41	1.31	1.10	0.90	0.67	0.37	0.00	0.99	0.91	0.77	0.63	0.47	0.26	0.00	0.58	0.54	0.45	0.37	0.27	0.15	0.00
	0.2 (medium)	1	0.56	0.52	0.44	0.36	0.27	0.15	0.00	0.39	0.37	0.31	0.25	0.19	0.10	0.00	0.23	0.21	0.18	0.15	0.11	0.06	0.00
		2	1.13	1.05	0.88	0.72	0.54	0.30	0.00	0.79	0.73	0.62	0.50	0.37	0.21	0.00	0.46	0.43	0.36	0.29	0.22	0.12	0.00
		3	1.69	1.57	1.33	1.08	0.80	0.45	0.00	1.18	1.10	0.93	0.75	0.56	0.31	0.00	0.69	0.64	0.54	0.44	0.33	0.18	0.00
		4	2.26	2.09	1.77	1.44	1.07	0.60	0.00	1.58	1.46	1.24	1.00	0.75	0.42	0.00	0.93	0.86	0.72	0.59	0.44	0.25	0.00
		5	2.82	2.62	2.21	1.79	1.34	0.75	0.00	1.97	1.83	1.55	1.26	0.94	0.52	0.00	1.16	1.07	0.91	0.74	0.55	0.31	0.00
	0.3 (large)	1	0.85	0.78	0.66	0.54	0.40	0.22	0.00	0.59	0.55	0.46	0.38	0.28	0.16	0.00	0.35	0.32	0.27	0.22	0.16	0.09	0.00
		2	1.69	1.57	1.33	1.08	0.80	0.45	0.00	1.18	1.10	0.93	0.75	0.56	0.31	0.00	0.69	0.64	0.54	0.44	0.33	0.18	0.00
		3	2.54	2.35	1.99	1.62	1.21	0.67	0.00	1.78	1.65	1.39	1.13	0.84	0.47	0.00	1.04	0.97	0.82	0.66	0.49	0.28	0.00
		4	3.39	3.14	2.65	2.15	1.61	0.90	0.00	2.37	2.19	1.85	1.51	1.12	0.63	0.00	1.39	1.29	1.09	0.88	0.66	0.37	0.00
		5	4.23	3.92	3.31	2.69	2.01	1.12	0.00	2.96	2.74	2.32	1.88	1.41	0.78	0.00	1.74	1.61	1.36	1.10	0.82	0.46	0.00
0.5 (balanced)	0.1 (small)	1	0.33	0.30	0.24	0.19	0.14	0.08	0.00	0.23	0.21	0.17	0.13	0.10	0.05	0.00	0.13	0.12	0.10	0.08	0.06	0.03	0.00
		2	0.66	0.60	0.48	0.38	0.28	0.15	0.00	0.46	0.42	0.34	0.27	0.19	0.11	0.00	0.27	0.24	0.20	0.16	0.11	0.06	0.00
		3	0.99	0.89	0.72	0.57	0.41	0.23	0.00	0.69	0.62	0.51	0.40	0.29	0.16	0.00	0.40	0.37	0.30	0.23	0.17	0.09	0.00
		4	1.32	1.19	0.97	0.76	0.55	0.30	0.00	0.92	0.83	0.68	0.53	0.39	0.21	0.00	0.54	0.49	0.40	0.31	0.23	0.12	0.00
		5	1.64	1.49	1.21	0.95	0.69	0.38	0.00	1.15	1.04	0.84	0.66	0.48	0.26	0.00	0.67	0.61	0.49	0.39	0.28	0.15	0.00
	0.2 (medium)	1	0.66	0.60	0.48	0.38	0.28	0.15	0.00	0.46	0.42	0.34	0.27	0.19	0.11	0.00	0.27	0.24	0.20	0.16	0.11	0.06	0.00
		2	1.32	1.19	0.97	0.76	0.55	0.30	0.00	0.92	0.83	0.68	0.53	0.39	0.21	0.00	0.54	0.49	0.40	0.31	0.23	0.12	0.00
		3	1.97	1.79	1.45	1.14	0.83	0.45	0.00	1.38	1.25	1.01	0.80	0.58	0.32	0.00	0.81	0.73	0.59	0.47	0.34	0.19	0.00
		4	2.63	2.38	1.93	1.52	1.11	0.60	0.00	1.84	1.66	1.35	1.06	0.77	0.42	0.00	1.08	0.98	0.79	0.62	0.45	0.25	0.00
		5	3.29	2.98	2.41	1.90	1.38	0.75	0.00	2.30	2.08	1.69	1.33	0.97	0.53	0.00	1.35	1.22	0.99	0.78	0.57	0.31	0.00
	0.3 (large)	1	0.99	0.89	0.72	0.57	0.41	0.23	0.00	0.69	0.62	0.51	0.40	0.29	0.16	0.00	0.40	0.37	0.30	0.23	0.17	0.09	0.00
		2	1.97	1.79	1.45	1.14	0.83	0.45	0.00	1.38	1.25	1.01	0.80	0.58	0.32	0.00	0.81	0.73	0.59	0.47	0.34	0.19	0.00
		3	2.96	2.68	2.17	1.71	1.24	0.68	0.00	2.07	1.87	1.52	1.20	0.87	0.48	0.00	1.21	1.10	0.89	0.70	0.51	0.28	0.00
		4	3.95	3.57	2.90	2.28	1.66	0.91	0.00	2.76	2.50	2.03	1.59	1.16	0.63	0.00	1.62	1.46	1.19	0.93	0.68	0.37	0.00
		5	4.93	4.46	3.62	2.85	2.07	1.13	0.00	3.45	3.12	2.53	1.99	1.45	0.79	0.00	2.02	1.83	1.48	1.17	0.85	0.46	0.00

[a] Of either of two constructs.

[b] ρ_{a_1,a_2} is the intra-educator, inter-construct correlation of the value added to two constructs that are combined to create a single vertical scale.

moderated performance categories can be affected by changes in the content of the tests as well as scores derived from a vertical scale. Therefore, while VMSS is an important part of solving the distortions, it is insufficient to resolve those problems alone.

Domain-referenced measurement

Domain-referenced measures of growth have also been advanced as a possible alternative to vertical scaling for the measurement of student progress (see Schulz et al., 2005). This also is a good beginning on resolving the distortions identified by Martineau (see above). Their approach divides the content of an assessment into multiple theoretically defined domains which can be followed across multiple years of assessment, providing measures of student growth on each of the content domains.

However, there are two difficulties with this approach. Because it depends on the theoretical definition of content domains, this approach may ignore the empirical conflation of multiple content domains into a single domain and/or the empirical need to separate out a single theoretically defined content domain into multiple domains. When content domains that should be conflated are not, what should be a single content domain is included more than once in an aggregate measure of student gains. When content domains that should be divided are not, all of the distortions identified earlier still apply. Therefore, while careful theoretical definition of multiple content domains is also an important part of the solution, it is also insufficient alone to resolve the distortions resulting from changing content over grades.

Stronger vertical content representation with adjacent-grade linking only

Martineau (in press) suggested that matrix sampling items from adjacent grades over many forms of the current grade level test such that all items from above and below grades are represented in the data set may be a useful approach. Martineau also suggested that only linking

adjacent grades (e.g. 4 and 5, 5 and 6, but not 4 and 6) may be a useful approach. He hypothesized that 100 percent representation of the adjacent grade-content in each grade's item-response data coupled with linking only adjacent grades would dramatically reduce the change in content representations in student scores across adjacent grades.

Further study (Martineau, January 2005) showed through simulation that while this approach did provide slight amelioration of the empirical construct shift, it was insufficient to allow for the use of such scales for the measurement of growth. Additional simulations (Martineau, 2005) investigated the impact of bootstrapped mean student performance, where multiple achievement estimates for each student were calculated based upon the linked items and random samples of the same numbers of items from the adjacent grade levels. This study showed that while the growth estimates derived from the bootstrapped estimates of student achievement were asymptotically unbiased, they were unacceptably imprecise, correlating less than 0.3 with simulated true growth scores.

Therefore, while stronger representation of the entire content of adjacent grade levels, and only linking adjacent grade level scales are also important parts of the solution, they are also insufficient alone to resolve the distortions in studies of growth resulting from changing content over grades.

Uninvestigated alternatives

There are several additional possibilities to ameliorate the effects of content changing over grades. They require either the modification of the test administration model or of the psychometric model used to calibrate student achievement.

Modified test administration models

Each of the modified test administration models depends upon testing at least a subset of the content more than once. It is important to recognize here that one can test previous content on a later test or test later content on an earlier test. The decision of which way to do this depends upon the interpretation of growth desired. If previous content is presented on later tests, the interpretation is the degree to which students have grown on previously assessed content. For low-performing populations of students, this may be the most appropriate approach. If future content is presented on current tests, the interpretation is the degree to which students have growth on content on which they are in general yet to be instructed. For students generally performing at or above grade level, this is may be the best approach.

Multiple test administrations per year. Current test administration models call for testing once per grade level, particularly under the No Child Left Behind Act of 2001 ("NCLB", 2002). If student achievement is measured on parallel forms of a test containing the same content, then the distortions attributable to shifting content disappear. There are two ways to test a full complement of parallel content in an appropriate manner.

First, the fourth grade assessment could be given both at the end of the third grade and the end of the fourth grade. The same could be done with the third grade assessment. This approach allows for measuring growth on third and/or fourth-grade content while eliminating the effects of shifting content. Second, multiple assessments of parallel content could be administered each year. This approach eliminates concerns about both shifting content and about holding educators accountable for differential summer losses/gains of different demographic groups of students.

Either approach would require at least doubling the amount of yearly testing. It is highly improbable that such a proposal would be acceptable to stakeholders in education. In addition, both approaches invite "gaming the system" by encouraging students to perform poorly on a given proportion of the assessment, and putting forth maximal effort on other portions of the assessments.

While the approach of multiple administrations per year is theoretically the best approach to entirely eliminating the effects of shifting content, the potential political, logistical, and ethical costs are too great for high-stakes implementation.

Supplement traditional assessment with sampled, out-of-level content. It may be unnecessary to double the amount of testing in order to salvage the validity of student growth measures. By increasing test lengths by a fraction (say one fourth or one third) it may be possible to obtain valid estimates of student gains. While precise estimates of student gains are unmistakably distorted when derived from traditional vertical scales, it is possible that less precise but more valid estimates of student growth may be obtained by sampling out-of-level content. This may be done in at least three ways.

First, by supplementing a traditional end of fourth-grade assessment with a representative sample of the end of fifth-grade content, a less precise, but more valid estimate of student growth on fifth grade content can be derived from the students' performance on the fifth grade content in fourth grade, and the student's performance on the end of fifth grade test. Given that the fourth-grade estimate of achievement on fifth-grade content is only sampled, it is less precise. However, this less precise estimate is likely to provide much more valuable information than the more precise, but distorted estimates derived from traditional vertical scales. One difficulty of this approach is that it would require the supplemented tests to be longer than traditional tests.

Another difficulty of this approach is whether it would fulfill the requirements of NCLB that accountability be based on grade-level content.

Second, supplementing traditional computer-based tests (CBTs) with out-of-level content would accomplish the same results. This would also require the supplemented CBTs to be longer than traditionally CBTs.

Third, supplementing a traditional computer-adaptive test (CAT) with out-of-level content would accomplish the same results with less cost in terms of testing time. Because all item responses add to the precision of estimates of student achievement, out-of-level item responses could be used to inform the stopping rule so that the supplemented CATs need not be much longer than a traditional CAT.

It is important to note in this section that it is unresolved what percentage of prior (or future) content is necessary to test out of level in order to obtain adequate estimates of student growth. This is a topic for future research.

Modified psychometric models

Martineau and Plank (submitted for publication) concluded that it may be necessary to increase the complexity of mainstream psychometric models to account for the complex nature of the shifting content across grades. Current psychometric models assume a single, stable construct across all grade levels that are linked together. Clearly, this cannot be defended as a reasonable assumption. Increasing the complexity of the psychometric models to reflect the complexity of the item response data is likely to face stiff opposition because of concerns about a "black box" that nobody can understand. However, it is short-sighted to insist on a simple psychometric model when simple models are insufficient to allow for the measurement of student growth.

Multidimensional Item Response Theory (see Reckase, 1989a; Reckase, 1989b; Reckase & McKinley, 1991; Tam, 1992) provides a psychometric model that is capable of representing the complexity of shifting content over grades. Recent advances in MIRT equating techniques (Min, 2003; Reckase & Martineau, 2004) have moved MIRT into the field of vertically scaling multidimensional test scores. While there is no efficient production model exists for one-grade MIRT analysis (let alone multi-grade vertical MIRT scaling), this appears to be the most promising alternative to traditional vertical scales without radically changing the test administration model.

Some difficulties exist with using MIRT to measure student growth. First, the identification of appropriate empirical dimensionality is a matter of discussion. The most common approaches (TESTFACT by Bock *et al.*, 2003; DIMTEST by Stout *et al.*, 1993) rely on dominant dimensionality to determine the appropriate number of dimensions. This is highly dependent upon the correlation among dimensions, reducing the estimated number of dimensions the higher the correlations. Martineau (2004, in press) showed that even with high but imperfect intercorrelations among dimensions, the distortions remain a considerable problem. Dominant-dimensionality-based approaches to the appropriate number of dimensions is likely to allow the distortions to continue while giving support to the use of fewer dimensions than necessary to eliminate the harmful effects of shifting content over grades.

Reckase *et al.* (2000) developed and demonstrated a vector-based approach to determining dimensionality that allows for highly correlated dimensions to each be considered important in and of themselves rather than together as a dominant dimension. This approach provides promise that using a more complex psychometric model (MIRT) in once-yearly testing may result in unbiased, valid, and relatively precise estimates of student growth.

However, even in using the vector approach to determining dimensionality, a difficulty remains. These analyses can be highly intensive and time-consuming. Particularly, identifying the meaning of dimensions is a very difficult task. One possible approach to alleviating this problem is to determine how far from true dimensionality one may be before estimates of student growth (based on composites of growth in all dimensions modeled) are too distorted to be useful for either high-stakes or research use. This would remove the need to name the dimensions in favor of being confident that the number of dimensions is close enough to provide accurate measures of growth.

Second, for a large-scale testing program, the cost of MIRT analyses is likely to be unreasonably high until efficient MIRT production models exist. The development of intuitive and user-friendly MIRT analysis tools will be very helpful in this process.

Discussion

Several alternatives to traditional vertical scaling were discussed in this paper. Those that have been studied (vertically moderated standard setting, domain-referenced measurement, and stronger content representation in vertical links with adjacent-grade linking only) have been found to be necessary, but insufficient to address the distortions in measures of student growth that are attributable to shifting content over grades in traditional vertical scales.

Alternatives that have not yet been studied are promising but difficult to implement. Some alternatives are modifications to traditional test administration models (multiple test administrations per year, out-of-level supplemented paper & pencil, computer based tests, and computer adaptive tests). These require significant political and logistical costs, including an increased amount of testing. The last alternative is a modification of the psychometric model to adequately represent to multidimensional complexity of grade-level-content-shifting vertical

scales. This alternative requires significant development work on a multidimensional psychometric production model, and introduces additional complexity that must be explained to stakeholders, but *does not* increase the amount of testing needed over that for a traditional vertical scale.

Michigan is investigating all of the above options for feasibility in overcoming the distortions in measures of student growth that are attributable to content-shifting vertical scales.

References

Angoff, W. H. (1971). Scales, Norms, and Equivalent Scores. In R. L. Thorndike (Ed.), *Educational Measurement* (second ed., pp. 508-600). Washington, D.C.: American Council on Education.

Ballou, D. (2002). Sizing Up Test Scores. *Education Next, 2*(2), 10-15.

Bereiter, C. (1963). Some Persisting Dilemmas in the Measurement of Change. In C. W. Harris (Ed.), *Problems in Measuring Change* (pp. 3-20). Madison, WI: University of Wisconsin Press.

Bock, R. D., Gibbons, R., Schilling, S. G., Muraki, E., Wilson, D. T., & Wood, R. (2003). TESTFACT 4. Lincolnwood, IL: Scientific Software International, Inc.

Braun, H. I. (2004, October 21-22). *Value-Added Modeling: What does "Due Diligence" Require?* Paper presented at the Conference on Value-Added Modeling: Issues with Theory and Application, College Park, MD.

Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998a). *Academic Productivity of Chicago Public Elementary Schools. Examining Productivity Series. A Technical Report*. Chicago, IL: Consortium on Chicago School Research.

Bryk, A. S., Thum, Y. M., Easton, J. Q., & Luppescu, S. (1998b). Assessing School Academic Productivity: the Case of Chicago School Reform. *Social Psychology of Education, 2*, 103-142.

Cizek, G. J. (2005a). Adapting Testing Technology to Serve Accountability Aims: The Case of Vertically Moderated Standard Setting. *Applied Measurement in Education, 18*(1), 1-10.

Cizek, G. J. (Ed.). (2005b). *Applied Measurement In Education* (Vol. 18).

Cliff, N. (1991). Ordinal Methods in the Assessment of Change. In L. M. Collins & J. L. Horn (Eds.), *Best Methods for the Analysis of Change: Recent Advances, Unanswered Questions, Future Directions* (pp. 34-46). Washington, D.C.: American Psychological Association.

Hoover, H. D. (1984). The Most Appropriate Scores for Measuring Educational Development in the Elementary Schools: GE's. *Educational Measurement: Issues and Practice, 3*(4), 8-14.

Huynh, H., & Schneider, C. (2005). Vertically Moderated Standards: Background, Assumptions, and Practices. *Applied Measurement in Education, 18*(1), 99-114.

Kingsbury, G. G., & McCall, M. (2005). *A Hybrid Model of School Success: Measuring Growth in the Context of Standards.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Lord, F. M. (1963). Elementary Models for Measuring Change. In C. W. Harris (Ed.), *Problems in Measuring Change* (pp. 21-38). Madison, WI: University of Wisconsin Press.

Martineau, J. A. (2004). *The Effects of Construct Shift on Growth and Accountability Models.* Unpublished Dissertation, Michigan State University, East Lansing.

Martineau, J. A. (2005). *Revisiting Alternatives to Vertical Scaling.* Paper presented at the March meeting of the Technical Advisory Committee of the Office of Educational Assessment & Accountability of the Michigan Department of Education, Lansing, MI.

Martineau, J. A. (in press). Distorting Value Added: The Use of Longitudinal, Vertically Scaled Student Achievement Data for Growth-Based Value-Added Accountability. *Journal of Educational and Behavioral Statistics*.

Martineau, J. A. (January 2005). *Undistorting Value Added? Investigating Alternatives to Traditional Vertical Scales for Growth-Based Value-Added Modeling.* Paper presented at the January meeting of the Technical Advisory Committee of the Office of Educational Assessment & Accountability of the Michigan Department of Education, Romulus, MI.

Martineau, J. A. (submitted for publication-a). Distorting Growth: The Use of Vertical Scales in Growth Models Using Complex Representations of Time. *Journal of Educational and Behavioral Statistics*.

Martineau, J. A. (submitted for publication-b). Distorting Growth: The Use of Vertical Scales in Growth Models Using Simple Representations of Time. *Journal of Educational and Behavioral Statistics*.

Martineau, J. A. (submitted for publication-c). Distortions in Growth Trajectories Based on Vertically Scaled Student Achievement Test Scores. *Journal of Educational and Behavioral Statistics*.

Martineau, J. A., & Plank, D. N. (submitted for publication). Promise, Betrayal, and Redemption: The Use of Vertical Scales in Growth-Based Value-Added Models. *Educational Evaluation and Policy Analysis*.

Martineau, J. A., Somers, J. G., & Dehlin, J. O. (submitted for publication). The Formation of Achievement Gaps on a Reading Achievement Test: Implications for Educational Policy. *Educational Evaluation and Policy Analysis*.

Min, K.-S. (2003). *The Impact of Scale Dilation on the Quality of the Linking of Multidimensional Item Response Theory Calibrations.* Unpublished Dissertation, Michigan State University, East Lansing, MI.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425, (2002).

Phillips, S. E., & Clarizio, H. F. (1988). Conflicting Growth Expectations Cannot Both be Real: a Rejoinder to Yen. *Educational Measurement: Issues and Practice, 7*(2), 18-19.

Popham, J. W. (1997). The Moth and the Flame: Student Learning as a Criterion of Instructional Competence. In J. Millman (Ed.), *Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure?* (pp. 264-274). Thousand Oaks, CA: Corwin Press.

Popham, W. J. (2005). *Discussant: Constructs and Methods in Measuring Growth.* Paper presented at the Annual Meeting of the American Educational Research Association, Montreal, Quebec, Canada.

Reckase, M. D. (1989a, August). *Controlling the Psychometric Snake: Or, How I Learned to Love Multidimensionality.* Paper presented at the Annual Meeting of the American Psychological Association, New Orleans, LA.

Reckase, M. D. (1989b, March). *Similarity of the Multidimensional Space Defined by Parallel Forms of a Mathematics Test.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Reckase, M. D. (2004). The Real World Is More Complicated Than We Would Like. *Journal of Educational and Behavioral Statistics, 29*(1), 117-120.

Reckase, M. D., & Martineau, J. A. (2004, October). *Growth as a Multidimensional Process.* Paper presented at the Annual Meeting of the Society for Multivariate Experimental Psychology, Naples, FL.

Reckase, M. D., Martineau, J. A., & Kim, J.-P. (2000, June). *A Vector Approach to Determining the Dimensionality of a Data Set.* Paper presented at the annual meeting of the Psychometric Society, Seattle, WA.

Reckase, M. D., & McKinley, R. L. (1991). The Discriminating Power of Items that Measure More than One Dimension. *Applied Psychological Measurement, 15*(4), 361-373.

Schaeffer, B. (2004). Districts Pilot Value-Added Assessment. *The School Administrator, 11*(12), 20-24.

Schmidt, W. H., & Houang, R. T. (2004, October 21-22). *The Role of Content in Value-Added Research.* Paper presented at the Conference on Value-Added Modeling: Issues with Theory and Application, College Park, MD.

Schulz, E. M., Lee, W.-C., & Mullen, K. (2005). A Domain-level Approach to Describing Growth in Achievement. *Journal of Educational Measurement, 42*(1), 1-26.

Seltzer, M. H., Frank, K. A., & Bryk, A. S. (1994). The Metric Matters: The Sensitivity of Conclusions About Growth in Student Achievement to Choice of Metric. *Educational Evaluation and Policy Analysis, 16*(1), 41-49.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. (1993). DIMTEST manual: Unpublished manuscript, University of Illinois, Urbana-Champaign.

Tam, S. S. (1992). *A Comparison of Methods for Adaptive Estimation of a Multidimensional Trait.* Unpublished Dissertation, Columbia University, New York.

Thum, Y. M. (2002). *Measuring Progress Towards a Goal: Estimating Teacher Productivity Using a Multivariate Multilevel Model for Value-Added Analysis*. Santa Monica, CA: Milken Family Foundation.

Wainer, H. (Ed.). (2004). *Journal of Educational and Behavioral Statistics* (Vol. 29).

Yen, W. M. (1985). Increasing Item Complexity: A Possible Cause of Scale Shrinkage for Unidimensional Item Response Theory. *Psychometrika, 50*(4), 399-410.

Yen, W. M. (1986). The Choice of Scale for Educational Measurement: An IRT Perspective. *Journal of Educational Measurement, 23*(4), 399-325.

Yen, W. M. (1988). Normative Growth Expectations Must be Realistic: a Response to Phillips and Clarizio. *Educational Measurement: Issues and Practice, 7*(2), 16-17.

Zwick, R. (1992). Statistical and Psychometric Issues in the Measurement of Educational Achievement Trends: Examples from the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*(2), 205-218.